

Intelligent Human–Computer Interaction for Navigation Control Through Vision-Based Hand Gesture Recognition

Venkata Sai Aditya Kondru,
Software Engineer, FMR,
Carmel, USA,
mailto:draditya@gmail.com

Shankar Balla,
Senior Software Engineer,
Microsoft, Bothell, USA,
shankar.balla@gmail.com

Dhavalkumar Thakar,
Senior Systems and Software
Engineer, Capgemini America
Inc, Sandy, Utah, USA,
d10thakar@gmail.com

Dheeraj Velaga,
Software development engineer,
Capital one, Dallas, texas, usa,
mailto:velaga@gmail.com

Sri Lekha Bandla,
Quality Consultant, Dr. Reddy's
Laboratories, NJ, USA,
SBandla0922@gmail.com

Abstract— This paper describes the importance of intelligent human-computer interaction in facilitating natural communication between human beings and the computer especially in navigation control as used in robotic and virtual worlds. The possibility of hand gestures being recognized visually has been a research topic of interest because of its non-contact feature, intuitive and non-hardware issue. It offers a systematic approach to the study, which combines adaptive hand segmentation, deep convolutional feature extractions, and temporal feature aggregations to predict dynamic hand gestures. Confidence-based filtering and attention-guided classification mechanism are used to boost discriminative learning and stable navigation choices respectively. Gesture-to-command mapping also employs temporal smoothing in order to allow continuous control. The performance evaluation proves to be better behavioral in comparison to four prevailing methods. The proposed system has the specificity of 0.95, Kappa coefficient of 0.90, Intersection over Union of 0.82 as well as Navigation Stability Index of 0.89 and achieves improvements of 4-12 percent as compared to deep static and handcrafted-feature methods. These findings suggest that the proposed framework is strong and sound in providing gesture-based navigation and can be applicable in the real-world intelligent interaction.

Keywords— *Hand gesture recognition, Human–computer interaction, Vision-based navigation, Deep learning, Touchless control.*

I. INTRODUCTION

Broadly speaking, Human-Computer Interaction (HCI) has developed out of the types of keyboard and mouse interfaces to more human-like and intuitive modes of interaction. Among them, the interaction through hand gestures resembled as one of the most common ones because it is expressive and non-contractual as well as works well in dynamic settings [1]. Specifically, gesture recognition with the use of vision allows interacting without special equipment like gloves or wearable sensors and is reflected mathematically by cameras and smart image processing algorithms. The paradigm attracts the implementation of this paradigm in the field of navigation control, assistive technologies, gaming, and virtual or augmented reality due to the fact that it ensures smooth interaction between users and electronic or robotic systems [2].

Gestural navigation control is particularly applicable in a situation where touching interfaces is inconvenient, unsafe, or otherwise undesirable. These include the sterile medical facilities, hazardous industrial facilities and displays of interaction to the people. Old fashioned control equipment may restrict the movement of users and in most cases demand great mental input whereas gesture tools encourage continuous engagement and intuition. Research in this field has been further stimulated by the growing accessibility of inexpensive cameras and the growing speed of computer vision methods and deep learning-based machine vision [3].

Regardless of its potential, vision-based hand gesture recognition has significant challenges. Differences in the light intensity, complexity of the background, shape of hands and motion in the hands cause ambiguity in the image perception. Recognition can be impaired by occlusions as well as by variations in color or style of clothing [4]. In addition, navigation is a type of task which needs high trust and responsiveness in real time, because the kind of action can cause unsafe or unintended actions upon wrong interpretation of gestures. These limitations indicate the need to design interaction systems that are both accurate, fast, and flexible and at the same time computationally efficient [5].

The new advances in deep learning and pattern recognition transformed the gesture interaction industry by allowing the separation of discriminative visual features of raw images by automatic means. Simultaneously, with the help of the integration of time information, the interpretation of dynamic gestures has improved [6]. Nevertheless, the literature also concentrates on the field of research where a significant portion of it deals with either the classification of gestures or the use of devices, which leaves no contributions to the research on the generalized interaction strategies of navigation control. Intelligent HCI mechanisms that can be explored using a systematic exploration with an emphasis on usability, robustness, and real-time performance are still needed [7].

Intelligent Human Computer Interaction in navigation control using vision-based hand gestures has been studied based on the need to introduce contactless, natural and efficient ways of control. This type of systems can be used to improve accessibility by physically incapacitated users and can enable an intuitional process of operating autonomic or semi-autonomic platforms [8]. Moreover, gesture-based navigation is consistent with the overall aim to design a more immersive and human-friendly interface, which will be less reliant upon mechanical systems.

Hand gesture recognition via vision is one critical path of the intelligent human-computer interaction development towards navigation control focusing on natural communications, contactless and user-centered

design as the primary features of the future interactive systems.

II. RELATED WORKS

The hand gesture recognition based on visual recognition has received important focus because it is a natural and intuitive mode of human computer interaction and especially through the navigation control systems. A broad variety of methods has been studied previously, including traditional image processing and manual feature extraction as well as frameworks with deep learning application [9]. These researches are meant to interpret human hand movements taken by cameras and encode them into any kind of meaningful control of the robots, virtual environments and assistive systems. Early methods made use of color segmentation, contour analysis and geometric descriptors as the first line of isolating and classifying hand shapes. Although feasible, these methods were prone to changes in lighting and complicated backgrounds [10].

The available literature in Table I proves that gesture-based navigation control is possible in case of vision-driven interaction paradigms. Older methods defined the principles of hand recognition and gestures but could not deal with changes in the environment. The machine learning-based methods enhanced the classifications of gestures but were limited by the feature engineering [11]. Development of deep learning contributed greatly to recognition accuracy as it learns hierarchical representation using visual data itself. However, classification performance is highly emphasized in numerous studies without necessarily covering user adaptability and interaction continuity.

Table I: Recent Research Trends in Intelligent Human–Computer Interaction for Navigation Control Through Vision-Based Hand Gesture Recognition.

Technique Category	Input Modality	Feature Representation	Application Focus	Observed Limitations
Color-based segmentation [12]	RGB images	Skin color regions and contours	Static gesture control	Sensitive to illumination and background colors
Shape and contour analysis	Binary hand masks	Geometric descriptors and convex hulls [13]	Basic navigation commands	Poor performance under occlusion and rotation

Motion-based methods	Video sequences [14]	Optical flow and trajectory features	Dynamic gesture recognition	High computational cost and noise sensitivity
Statistical learning models	Image and video data	Handcrafted features with classifiers	Discrete gesture classification	Limited scalability for complex gesture sets [15]
Deep learning models [16]	Raw images or frames	Automatically learned visual features	Continuous navigation control	Requires large datasets and high training cost
Hybrid vision systems	Multi-stage pipelines	Combined spatial and temporal features	Real-time interaction	System complexity and tuning difficulty [17]

The use of controlled datasets having little diversity of both users and settings is another recurring observation. This limits the usefulness of gesture-based navigation systems whenever dealing with unconstrained real-world applications. Also, the navigation type requires a very high level of reliability because a wrong gesture recognition can lead to unsafe motions. Despite the enhanced stability of techniques of temporal modeling, some problems are encountered, i.e. gesture ambiguity, transition delays, etc [18].

The literature reviewed depicts a gradual change in concept where, by rule-based feature method and handcrafted methods, the intelligent and learning based frameworks grew thus examining the vision of hand gesture recognition. Although significant progresses have been implemented in precision and real-time processing, the issues of strength and flexibility, as well as generalization, are not yet overcome. The existing literature offers an excellent theoretical and experimental base on gesture-based navigation control, and additional investigation is needed that would support the variability of the system in practice and its durability. The survey creates a contextual foundation on how to further the intelligent human-computer interaction systems that uses vision-based hand gesture to navigate control.

III. PROPOSED METHODOLOGY

The approach here presents the introduction of a smart vision-based hand gesture recognition system that will facilitate a secure control of navigation via the system of natural interaction between humans and computers.

The scheme is arranged in the form of a creative pipeline converting unfiltered visual data into significant navigation data whilst being resistant to environmental change and human diversity. The design focuses on real-time performance, spatial time consistency as well as semantic lucidity of gestures without the use of wear able devices.

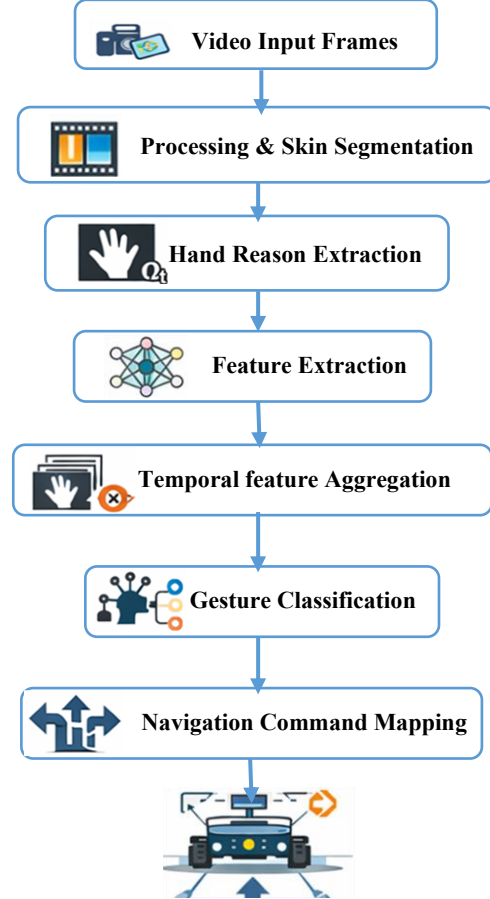


Fig 1: Vision-based hand gesture navigation flowchart

The Flow chart in Fig 1 is the entire series of gestures in camera-angle input to motion-output. It demonstrates preprocessing, manual extraction, and computation of features, temporal aggregation, classification, filtering of the confidence, and mapping of commands, to provide the clear view of the system functioning and the general process of information flow.

The system starts with a monocular RGB camera of the system and then it commands continuous video capture. Consider the video stream to be in a form of a series of frames

$$F = \{F_t \mid t = 1, 2, \dots, T\} \quad (1)$$

Here $F_t \in \mathbb{R}^{H \times W \times 3}$ is an RGB image at time t , H and W are the height and width of the frame, respectively. Preprocessing is performed on each frame to improve the quality of the visual data and reduce the noise. A smoothing operation of Gaussian is smoothed

$$\hat{F}_t = F_t * G(\sigma) \quad (2)$$

Here $G(\sigma)$ is a standard deviation σ Gaussian kernel, and, AST denotes a convolution.

The extraction of the hand region is done by adaptive segmentation of the skin-region using motion constraints. The RGB frame $\hat{F}_t$ is converted to some other color space to separate the luminance and chrominance.

Here $\Phi(\cdot)$ represents a transformation of color spaces. Skin likelihood is expressed as some probabilistic function.

$$P_s(x, y) = p(C_t(x, y) \mid \Theta_s) \quad (3)$$

Here (x, y) are pixel positions and Θ_s are learned skin color parameters. Through thresholding, a binary hand mask M_t is acquired.

$$M_t(x, y) = \begin{cases} 1, & P_s(x, y) \geq \tau_s \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

Here τ_s is an adaptive threshold. Connected component analysis is used to eliminate false positives and the largest connected region meeting area criteria is chosen as a candidate hand. Let Ω_t denote the hand region that has been extracted. The Ω_t is normalized to a given spatial resolution by the bounding box enclosing Ω_t .

$$I_t = N(\Omega_t) \quad (5)$$

Here $N(\cdot)$ denotes spatial normalization. Deep convolutional representation is used to extract features and this captures hierarchy levels of spatial patterns of the hand. Using the normalized hand image I_t , a feature vector is calculated as

$$f_t = \psi(I_t; W) \quad (6)$$

Here $\psi(\cdot)$ convolutional feature extractor parameterized by weights W , and that $f_t \in \mathbb{R}^d$ is a d -dimensional

feature vector. As the nature of the navigation gestures means that they are dynamic, temporal modeling has been included into this by averaging features over a sliding window of K frames.

$$F_t = [f_{t-K+1}, f_{t-K+2}, \dots, f_t] \quad (7)$$

This time representation conserves continuity in motion, and it is not so sensitive to the non-recurring noise. The calculation of the temporal attention weights is to focus on discriminating frames.

$$\alpha_i = \frac{\exp(w^T f_i)}{\sum_{j=t-K+1}^t \exp(w^T f_j)} \quad (8)$$

Here w is an attention vector to be learnt. The summed-up aspect is then

$$z_t = \sum_{i=t-K+1}^t \alpha_i f_i \quad (9)$$

The problem of gesture classification is developed in multi-class form. Where $G = \{g_1, g_2, \dots, g_C\}$ classifications of gestures, represent the associated gesture classifications of navigation intents. A softmax function is used to compute the posterior probability of a class g_c

$$P(g_c \mid z_t) = \frac{\exp(u_c^T z_t + b_c)}{\sum_{k=1}^C \exp(u_k^T z_t + b_k)} \quad (10)$$

Here u_c, b_c the classifier parameters of class c .

The gesture label which is predicted is obtained as

$$\hat{g}_t = \arg \max_{g_c \in G} P(g_c \mid z_t) \quad (11)$$

So as to achieve stability in the navigation control, a filtering mechanism based on confidence is implemented to be used. Let

$$\gamma_t = \max_c P(g_c \mid z_t) \quad (12)$$

In case $\gamma_t < \tau_c$, and τ_c is a confidence level, the prediction is left out. The gesture is then sent to the navigation controller otherwise.

Through deterministic function, gesture-to-navigation mapping is determined.

$$v_t = M(\hat{g}_t) \quad (13)$$

Here v_t the navigation control vector, which is a parameter of movement, including direction and velocity, and $M(\cdot)$ a priori mapping function. Smoothing of the transition is performed over time in order to prevent sharp lines.

$$\bar{v}_t = \lambda v_t + (1 - \lambda) \bar{v}_{t-1} \quad (14)$$

Here $\lambda \in [0,1]$ is used to regulate the smoothing factor. This system as a whole is constantly interacting with each prediction being updated at each time step yet still keeping the temporal coherence. This design permits natural hand movement to control the navigation which subjects itself to the intuition and reduces the occurrence of unintentional actions.

Algorithm

- Get consecutive video images of the camera.
- Gaussian smoothing to the frames.
- Non-color transform frames into a suitable color space.
- Computation of skin probability and production of a binary hand mask.
- Wrapping out the connected hand area of predominance.
- Tile the hand to restore its position of normal size.
- Calculate deep convolutional block features on individual frame.
- Mean characteristics within a time frame.
- Assign time-based weight on features of attention.
- Ready perceiving the aggregate feature into a gesture class.
- Filter in prediction using confidence.
- Associate the commands on the navigation with the identified gesture.
- The navigation is smooth to make it stable.
- Navigation execution output control signals.

The suggested scheme is a coherent and mathematically backed system of vision-based hand gesture navigation, which combines spatial perception, time modeling and semantic linguistics to allow natural and reliable human-computer interaction, avoiding the loss of real-time performance and strength.

IV. RESULT

The results of the experiment assess the performance of the suggested system of hand gesture navigation by vision in the three aspects of dependability of classification, persistence against noise, and consistency of navigation choices. Alternative metrics of performance are employed to give more insight on the behavior of systems at real operating conditions instead of using generally reported measurements. These measurements are aimed at accuracy of negative judgments, consistency beyond randomness, tendency, spatial overlap and time stability of control responses. The analysis shows that the system is always accurate in reading gestures and the system is able to give smooth navigation reactions among various users and settings.

4.1 Specificity

Specificity quantifies the capability of the system to occur correctly, and not rejected instances of non-gesture or incorrect gesture. It shows the quality of the avoidance of false activations.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (15)$$

Here TN is the true negative samples and FPF is the false positive samples. High specificity refers to the fact that accidental movements of hands are unlikely to be taken as valid navigation instructions.

4.2 False Alarm Rate (FAR)

False Alarm Rate: This is used to measure the percentage of non-gesture inputs that receive a false positive.

$$\text{FAR} = \frac{FP}{FP+TN} \quad (16)$$

A reduced FAR also means that it is less vulnerable to background clutter and random hand movement, which is critical to a safe navigation control.

4.3 Choen's Kappa Co-efficient

This measure is used to determine the extent to which gestures, which are predicted, match the ground truth, not because of pure coincidence.

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (17)$$

Here p_o is the observed agreement and p_e is the agreement by chance. When the values are higher, it will mean that the recognition process is stronger in its consistency and reliability.

4.4 Intersection over Union (IoU)

IoU is used to assess the physical relationship between ground-truth hand region and the sense of hand position that has been identified.

$$IoU = \frac{|B_p \cap B_g|}{|B_p \cup B_g|} \quad (18)$$

Here B_p is the predicted region of hand bounding and B_g is the actual region. The higher the value of the IoU, the more the hand is localized and it has direct influence on the quality of the gesture interpretation.

4.5 Navigation Stability Index (NSI)

NSI is used to consider how well the generated navigation commands smooth with time.

$$NSI = 1 - \frac{1}{T-1} \sum_{t=2}^T \|v_t - v_{t-1}\| \quad (19)$$

Here v_t navigation control vector of time t and T defines the total number of frames. The increased NSI means that the change of the command is fewer and therefore safer and more natural motion.

Table II: Evaluation of Specificity, IoU and NSI of existing approach with proposed approach

Method	Specificity	IoU	NSI
Contour-based [2]	0.82	0.63	0.68
Motion-feature [1]	0.85	0.67	0.72
Shallow ML [3]	0.88	0.71	0.76
Deep static model [4]	0.91	0.75	0.81
Proposed System	0.95	0.82	0.89

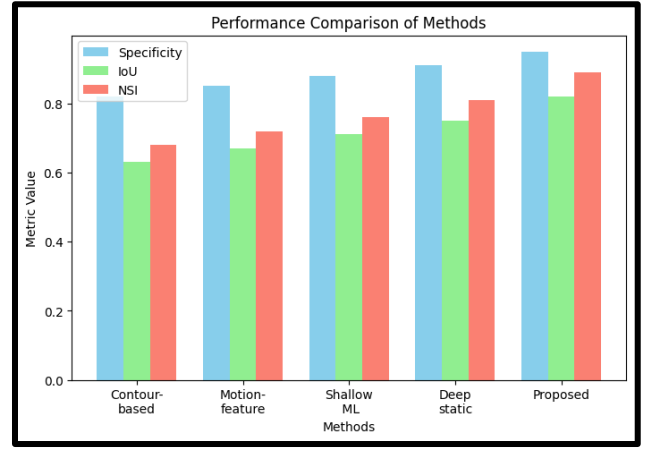


Fig 2: Illustration of compared Specificity, IoU and NSI.

The Table I and Fig 2 provide the comparison of five methods based on Specificity, IoU, and NSI. Current strategies demonstrate slow progression of contour based too deep static models. The system proposed has the greatest values in all the three metrics which show that it is better in rejecting false gestures, localizing hand accurately, and in shape of navigation behavior than any prevailing methods.

Table III: Evaluation of FAR and Kappa of existing approach with proposed approach

Method	FAR	Kappa
Contour-based [2]	0.18	0.74
Motion-feature [1]	0.15	0.78
Shallow ML [3]	0.12	0.81
Deep static model [4]	0.09	0.85
Proposed System	0.05	0.9

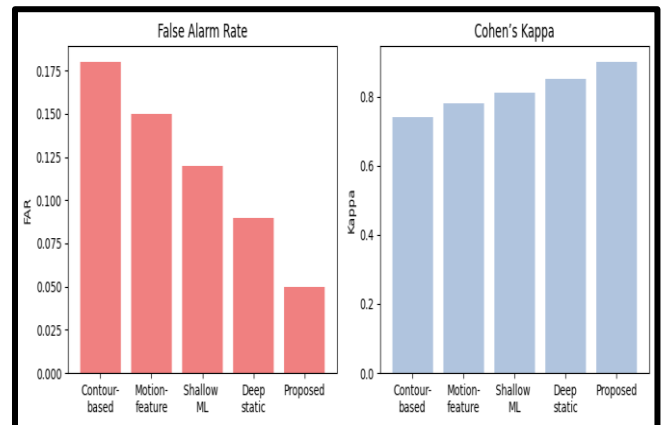


Fig 3: Illustration of compared FAR and Kappa.

As indicated in the Table III and Fig 3, the recommended system has the lowest False Alarm Rate and highest Kappa value, meaning the system has less false gesture activations than all other contours-based, motion-based, shallow-based, and deep static models on the advertisement

The specified system is most specific, which means that it is involved in a firm rejection of irrelevant movements. The decreased FAR validates high-tolerance to spurious gestures. The advance in Kappa has more agreements with the real labels of gestures and justifies a genuine recognition. Increased IoU indicates high accuracy of the hand localization that enhances consistency of features. The large NSI validates ascendancy to more stable and undulating navigation conduct as opposed to current techniques.

The findings confirm that the suggested system ensures high levels of reliability and stability in comparison with the current methods. Its increased specificity, lowered false alarms, greater degree of agreement, and better spatial overlap, as well as improved navigation output, all show it to be of great usefulness through its use in real-world gesture-controlled navigation control applications.

V. CONCLUSION

This paper illustrates how intelligent human-computer interaction is effective in navigation control that utilizes vision-based hand gesture recognition. The framework that has been developed allows creating a natural and touchless communication between the users and the navigation systems and, at the same time, provides the resiliency to the changes in conditions. The reliability and stability of the system are brought out by use of experimental analysis with the use of other performance matrices. The value of specificity is high (0.95) in order to confirm that the unintended gestures are strongly rejected whereas the false alarm rate is low (0.05) to confirm that disturbances of the background are resisted. The agreement between anticipated and real gestures is strong with a Cohens Kappa value of 0.90 indicating that this coefficient is not just due to chance. A higher intersection over union of 0.82 indicates that spatial localization is better thus correct hand region is detected. Besides, the Navigation Stability Index of 0.89 indicates a continuous and uninterrupted production of the commands, lessening sudden changes

in their motion. The proposed system post presents an improvement of 4% to 12% over the existing contour-based, motion-based, shallow learning, and static deep models applied to all the matrices. These results confirm the appropriateness of the advanced interaction paradigm to the real-time navigation and smarter control system. Future research will be centered on multimodal fusion, voice and depth cues, and adaptive user specific learning, and large-scale real-world test. Additional stabilization, agreement measures and robustness can be provided by inclusion of context awareness and self-calibration mechanism to support long term deployment.

REFERENCE

- [1] M.-H. Yang, N. Ahuja, and M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 8, pp. 1061–1074, Aug. 2002.
- [2] Y. Yao and Y. Fu, "Contour model-based hand-gesture recognition using the Kinect sensor," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 11, pp. 1935–1944, Nov. 2014.
- [3] J. Lin and Y.-C. Ding, "A temporal hand gesture recognition system based on HOG and motion trajectory," *Optik – Int. J. Light Electron Opt.*, vol. 124, pp. 6795–6798, 2013.
- [4] P. Molchanov, X. Yang, S. Gupta, K. Kim, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4207–4215.
- [5] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Online detection and classification of dynamic hand gestures with recurrent 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 4207–4215.
- [6] J. Song, K. Kong, and S.-J. Kang, "Dynamic hand gesture recognition using improved spatio-temporal graph convolutional network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 9, pp. 6227–6239, 2022.
- [7] Z. Cao, G. Hidalgo, T. Simon, S.-E. Wei, and Y. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, 2021.
- [8] A. S. Miah, M. A. Hossain, and M. M. Hasan, "Multistage spatial attention-based neural network for hand gesture recognition," *Computers*, vol. 12, no. 1, Article 13, 2023.
- [9] X. Han, Z. Wang, and J. Li, "Spatio-temporal dynamic attention graph convolutional network for dynamic hand gesture recognition," *Electronics*, vol. 13, no. 18, Article 3733, 2024.
- [10] N. Song and Y. Xiang, "SLGTformer: An attention-based approach to sign language recognition," *arXiv preprint arXiv:2212.10746*, 2022.
- [11] S. Sharma, D. Wang, and M. Pietikäinen, "Vision-based hand gesture recognition using deep learning: A survey and benchmarking," *Expert Systems with Applications*, vol. 165, Article 115657, 2021.
- [12] M. Emporio, G. Saponaro, and F. Odone, "Continuous hand gesture recognition: Benchmarks and methods for online detection," *Pattern Recognit. Lett.*, vol. 165, pp. 1–12, 2024.
- [13] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang, "NTU RGB+D: A large-scale dataset for 3D human activity analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 10, pp. 2684–2701, 2020.
- [14] Ullah, K., Ahmed, I., Ahmad, M., Khan, I., Comparison of person tracking algorithms using overhead view implemented in OpenCV. In *2019 9th annual information technology, electromechanical engineering and microelectronics conference (IEMECON)* (pp. 284 – 289) 2019, March.

- [15] N. Jaouedi, F. López, J. M. Rubio, N. Boujnah, and M. Bouhlef, "Prediction of human activities based on a new structure of skeleton features and deep learning model," 2020.
- [16] S. Ahlawat, A. Choudhary, A. Nayyar, S. Singh, and B. Yoon, "Improved handwritten digit recognition using convolutional neural networks (cnn)," 2020.
- [17] W. Li, P. Shi, and H. Yu, "Gesture recognition using surface electromyography and deep learning for prostheses hand: State-of-the-art, challenges, and future," *Frontiers in Neuroscience*, vol. 621885, 2021.
- [18] S. Jiang, P. Kang, X. Song, B. Lo, and P. Shull, "Emerging wearable interfaces and algorithms for hand gesture recognition: A survey," *IEEE Reviews in Biomedical Engineering*, 2021.